中国科学: 生命科学　　2019 年　第 49 卷　第 4 期: 445 ~ 455
SCIENTIA SINICA Vitae　　lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述　　中国演化生物学进展与展望专辑

群体基因组学方法: 从经典统计学到有监督学习

施怿[1,3], 李海鹏[1,2*]

1. 中国科学院上海生命科学研究院, 中国科学院上海营养与健康研究所,中国科学院马普学会计算生物学伙伴研究所,中国科学院计算生物学重点实验室, 上海 200031;
2. 中国科学院动物进化与遗传前沿交叉卓越创新中心, 昆明 650223;
3. 中国科学院大学, 北京 100049
* 联系人, E-mail: lihaipeng@picb.ac.cn

收稿日期: 2018-12-20; 接受日期: 2019-01-31; 网络版发表日期: 2019-03-21
中国科学院战略性先导科技专项(批准号: XDB13040800)和国家自然科学基金(批准号: 91531306, 91731304)资助

摘要　群体遗传学的一个主要研究目标是理解突变、自然选择、遗传漂变、群体结构和数量变化等进化力量如何共同影响基因组中的遗传变异. 通过分析DNA序列多态数据, 可以推测曾经作用于基因组的各种力量, 进而探讨生物演化的过程. 近年来, 随着第二代DNA测序技术的快速革新, 群体遗传学进入了基因组学时代, 相关的方法在不断发展, 并可将群体基因组学方法分为经典统计学方法和新兴的机器学习方法. 前者包括经典群体遗传学统计量、单一统计量或多统计量联合检测自然选择、群体历史与自然选择的联合估计以及基于溯祖树和祖先重组图的方法. 后者主要基于有监督学习, 为群体基因组时代的大数据分析带来了全新范式. 本文从理论基础出发, 全面回顾了群体基因组学方法发展变化的历程, 着重介绍了该领域的最新进展, 并就未来的发展方向进行了展望.

关键词　群体基因组学, 自然选择, 重组率, 经典统计学, 有监督学习

现代生物演化的观点经历了近两个世纪的发展历程. 1859年, 查尔斯•达尔文(Charles Darwin)《物种起源》一书中提出了自然选择学说[1]. 自1856年至1863年间, 格里哥•孟德尔(Gregor Mendel)完成了著名的豌豆杂交实验, 总结出孟德尔遗传定律, 并因此被后人看作现代遗传学的奠基人. 1942年, 朱利安•赫胥黎(Julian Huxley)将达尔文的自然选择学说与孟德尔的遗传学观点相结合, 提出了现代综合论[2]. 从第一次世界大战结束到20世纪50年代, 数学家罗纳德•费雪(Ronald Fisher)、生物学家莱特(Sewall Wright)与霍尔登(J. B. S. Haldane)发表的一系列论文率先将数学、尤其是统计学, 引入现代演化生物学研究中, 由此开创了群体遗传学这一交叉学科[3]. 1968年, 木村资生(Kimura)[4]提出了分子进化的中性理论, 认为遗传漂变等随机事件才是生物演化的主要动力, 看似与达尔文的自然选择学说相矛盾. 虽然从那时起, 自然选择学说和中性理论的支持者持续争论, 但后者其实为前者留下了足够的空间, 甚至可以看作是对前者的补充完善. 截至目前, 大多数生物学家都认同这样的观点, 即负选择是普遍存在的, 而正选择相对罕见, 同时认为正选择在

引用格式: 施怿, 李海鹏. 群体基因组学方法: 从经典统计学到有监督学习. 中国科学: 生命科学, 2019, 49: 445–455
Shi Y, Li H P. Population genomics: From classical statistics to supervised learning (in Chinese). Sci Sin Vitae, 2019, 49: 445–455, doi: 10.1360/N052018-00207

　

生物适应环境的演化过程中发挥重要作用[5].

进入21世纪，DNA测序技术的迅猛发展将整个生命科学研究领域带入了基因组学时代. 以大规模群体基因组学数据为起点，通过分析基因组中的遗传变异，研究曾经作用于基因组的自然选择、群体结构和数量变化历史等力量，从而探讨种群演化过程的群体基因组学应运而生. 数据是群体基因组学分析的起点. 2016年, Voight等人[6]利用国际单倍体计划(the International HapMap Project)[7]获得的人类基因组单核苷酸多样性(single nucleotide polymorphism, SNP)数据在全基因组水平上进行扫描并绘制出了人类基因组中受正选择影响的位点图谱. 2015年, 联合了来自中、日、英、美、肯尼亚等国多家机构的多学科研究团队共同完成的千人基因组计划(the 1000 Genomes Project)发表了第三阶段的最终成果[8,9]，同年，英国万人基因组计划(UK10K)也发表了大规模的群体基因组测序成果[10,11]. 这些项目为科学界和社会公众贡献了巨大的测序数据集. 除了测序得到的真实数据集，计算机模拟软件也可以为群体基因组学研究提供大量的模拟数据(*in silico* data sets). Hoban等人[12]综述了近年来的常用模拟软件，这些软件为理解难以分析预测的复杂演化模型提供了可能，也是理论群体基因组学的重要数据来源.

群体基因组学的一个重要研究方向是现代人群以及家养动植物的演化过程. 复旦大学金力院士的研究团队[13]通过分析Y染色体的数据，证明了现代东亚人群也起源于非洲. 中国科学院昆明动物所的He等人[14]采用多种分析方法、结合实验验证，探究了西藏人群适应高海拔缺氧环境的演化生物学机制. Marciniak和Perry[15]综述了运用古人类基因组研究现代人类适应性演化历史的最新进展. Wang等人[16]通过分析家犬、豺以及灰狼的基因组研究了家犬的起源. 相对于在人群以及家养动植物中开展的宏观研究，在细胞的微观水平上，一个重要的研究方向是癌症的演化过程(cancer evolution). 这一研究方向试图用演化生物学的观点来研究肿瘤的发生和发展[17,18]，具体研究内容有肿瘤异质性、癌细胞抗药性的起源、肿瘤遗传学、肿瘤表观遗传学以及癌细胞在肿瘤微环境层面受到的选择压力. 台湾大学的Wang和中山大学的Chen等人[19]进一步探讨了推动肿瘤发生发展的选择力量是目前学界主流认为的达尔文选择(Darwinian selection)或正选择，还是非达尔文选择(non-Darwinian selection)，在细胞水平上再次提出了自然选择学说和中性理论的争论.

工欲善其事, 必先利其器. 群体基因组学研究离不开有效的工具和方法，尤其是当基因组数据在测序、注释和比对过程中都不可避免地存在一些错误的时候[20]. 以自然选择学说和中性理论为基础，结合统计学的思想，群体基因组学利用从序列数据中获得的遗传变异信息，计算出一系列统计量来描述群体的特征，并进一步推测其在进化过程中受到的各种进化力量的作用. 例如, 在检测自然选择时, 通常以中性模型为零假设，衡量真实基因组情况与中性模型假设的偏离度从而判断该基因组区域是否受到自然选择作用. 因此本文不但简要总结了群体基因组学方法的发展历程，侧重于介绍最新进展，而且就未来的发展方向进行了展望.

1　经典统计量和自然选择检测方法

1.1　经典群体遗传学统计量

在进行群体遗传学分析时，由于真实群体的情况往往过于复杂，研究中需要基于一些理论假设对真实情况进行简化，获得理想化的群体模型. 所有这些模型和统计量中，Wright[21,22]的两篇论文中提出的有效群体大小N_e的概念是多种统计量和假设检验的重要基础.

Watterson[23]在理想化的种群模型中引入突变率μ但不考虑重组和迁移, 提出了群体突变率的估计值θ_W:

$$\theta_W = K / a_n,$$

$$a_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n-1},$$

其中K表示分离位点(segregating sites)的数量, n表示样本量. θ的估计值还有Tajima[24]提出的θ_π、Fu和Li[25]提出的θ_{FL}、Fay和Wu[26]提出的θ_H以及Zeng等人[27]提出的θ_L. 与θ_W的思想相类似, 这些方法都可以使用能从真实DNA序列数据中直接获得的数值来估计θ. 基于系统发育树的信息, Fu[28]也提出了新的θ估计值, 其精确度超过了已有的方法，并非常接近理论上的最优估值[29].

群体固定指数F_{ST}是Wright[30]提出的F统计量(Wright’s F-statistics)的一种形式，用于衡量亚群体

(subpopulation, S)和总群体(total population, T)之间的差异. F_{ST}的取值范围从0到1, 较小的F_{ST}值意味着群体的分歧较小, 而F_{ST}较大时则意味着群体的分歧较大. 随着DNA测序技术的发展, 目前更为常用的方法是Weir和Cockerham[31]提出的F_{ST}无偏估计量(Weir & Cockerham's unbiased F_{ST}). 该方法假设需要对i个亚群体进行分析(i=1, …, r), 利用从DNA序列信息中统计出的各个亚群体基因频率数据和各个亚群体的大小计算出可靠的F_{ST}估计量.

在群体水平上, 重组以一定的频率发生. 用r表示重组率, 则群体重组率可表示为: $\rho=4N_er$. Hudson和Kaplan[32]提出了最早的群体重组率估计方法. 该方法以R表示在样本群体历史中真实发生过的重组事件数. 虽然群体重组率ρ的公式与群体遗传多态性参数θ的公式非常相似, R却不能类比于K, 因为仅根据当前样本序列, 无从直接统计出该序列上曾经发生过的重组事件数. 但根据样本序列, 可使用最大简约法推断出重组事件数, 真实发生过的重组事件数的下界R_M也是可求的. 由此就可通过数值模拟的方法, 在给定θ和K的情况下, 求出R_M的均值、方差和分布, 从而估计重组率.

1.2　单一统计量或多统计量联合检测自然选择

自然选择是演化生物学研究中的一个重要研究内容. 在研究群体基因组学时, 通常将自然选择分为正选择(positive selection)、负选择(negative selection)和平衡选择(balancing selection)三种类型[33]. 正选择发生时, 受到选择的等位基因往往在当时的环境条件下更有利于物种的生存和繁衍. 伴随着该等位基因频率在群体中的迅速升高, 与该等位基因连锁的另一个"邻居"等位基因频率随之升高, 这一现象称为遗传搭车效应(hitchhiking). 负选择与正选择相反, 受到选择的等位基因对物种生存有害而被迅速淘汰出群体, 因此负选择又称为纯化选择(purifying selection). 与正选择和负选择不同的是, 平衡选择在多数情况下一般会倾向于维持群体的遗传多样性.

以遗传多样性变化为依据、构建单一统计量来检测自然选择的经典统计学方法可具体划分为四大类[34~36]: 基于比率的方法、基于基因频率的方法、基于连锁不平衡和选择扫荡的方法、基于种群分化的方法, 这些方法能够从宏观或微观的层面检测自然选择. 例如, 中国科学院北京基因组所的陈华研究组所开发的hmmSweep[37]和XP-CLR(the cross-population composite likelihood ratio method)[38], 以及Yi等人[39]的PBS(population branch statistic)分别使用基于连锁不平衡、选择扫荡和种群分化的方法检测自然选择, 均取得了积极的研究进展. 根据Vitti等人[34]和Akey[40]的总结, 加上我们的补充完善, 在表1中列出了一些具有代表性的具体方法.

上述方法各有优劣, 也都曾在不同尺度的群体基因组学研究中取得一定的成果. 在宏观的多物种尺度上, 经典的K_a/K_s比率方法可用于衡量某个编码基因在不同物种间突变速率的差异, 进而探讨这些物种的演化历史[51], 但该方法局限于编码蛋白质的基因组区域. 鉴别加速进化区的方法打破了这一限制, Pollard等人[45]以黑猩猩基因组为参照, 发现在7~19孕周的胎儿新皮层特定神经元中特异表达的RNA基因*HAR1F*经历了加速进化. 相比于基于比率的方法, 基于连锁不平衡和选择扫荡的方法更擅长在同一物种群体内或群体间的尺度上寻找选择信号. 在微观的单物种尺度上, XP-CLR方法可用于研究经济作物的驯化(domestication)过程, Hufford等人[52]和北京遗传发育所的Zhou等人[53]分别以玉米、大豆为对象, 使用XP-CLR方法寻找这些经济作物基因组上的选择信号, 从而为育种工作提供指导.

随着检测自然选择的统计量越来越多, 研究人员开始关注这些方法的特异性. 过去近40年的多项研究均表明, 使用单一的统计量进行群体遗传学估计时, 可能会得到有偏差的结果. 例如, 自然选择和群体历史就很有可能在基因组上留下相似的印记, 在进行相关估计时(如检测正选择时)如何对两者进行区分是一个关键的问题. 多篇综述[34,54,55]都提到了解决这一问题的思路: 将多个统计量联合起来, 共同检测自然选择.

中山大学的Zeng等人[27]提出了DH检验, 首次联合了Tajima's D和Fay & Wu's H来检测自然选择. Tajima's D在检测正选择时更倾向于频谱中的低频突变信号, 也因此更容易受到其他因素的干扰. 而Fay & Wu's H是基于高频突变的检验. 结果发现, DH检验对于正选择的检测特异性较高, 也更不容易受到其他因素的影响. 在检测新近发生的正选择以及达到高频率而尚未固定的有利突变时具有比Tajima's D检验和Fay

表 1 检测自然选择的经典统计学方法

Table 1 Classical statistics for detecting natural selection

所属类别	思想描述	代表性方法
基于比率的方法	假设同义突变在选择上是中性的, 其速率可以反映演化的背景速率. 如果非同义突变率与同义突变率显著不同, 就可以说明有自然选择的作用.	K_a/K_s(也被称为d_N/d_S或ω)[41,42], M.K.检验(McDonald-Kreitman test, MKT)[43]
	在一个谱系中经历了加速进化、而在另一个亲缘关系较近的谱系中保守的基因组区域可能受到了选择.	鉴别HARs(human accelerated regions)[44,45]
基于频率的方法	新出现的等位基因初始频率很低, 而一旦某个等位基因受到自然选择的作用, 其频率及其附近连锁基因的频率就会迅速升高(正选择)或者迅速从基因组中消失(负选择).	Tajima's D[24], Fu & Li's D[25], Fay & Wu's H[26]
基于连锁不平衡和选择扫荡的方法	选择扫荡(selective sweep)使遗传多样性降低, 并使特定等位基因及其邻近基因在群体中所占的比例升高, 这些基因共同定义了一个单倍型(haplotype), 该单倍型将在种群中持续占据主导地位, 直到重组(recombination)打破基因连锁关系.	Kim & Stephan[46], hmmSweep[37], EHH (extended haplotype homozygosity)[47], iHS (integrated haplotype score)[6], LDD (linkage disequilibrium decay)[48], XP-CLR[38]
基于种群分化的方法	如果某个基因在一个种群中受到选择, 而在另一个当中没有受到选择, 该基因就会在两个种群之间呈现出很大的频率差异, 这种差异会比未受到选择的基因表现出的频率差异更加显著.	F_{ST}, L.K.检验(Lewontin-Krakauer test, LKT)[49], LSBL(locus-specific branch length)[50], PBS[39]

& Wu's H检验方法更高的统计学功效. Zeng等人[56]还进一步尝试将EW检验(Ewens-Watterson test)[57]联合进来, 提出了联合Fay and Wu's H和EW的HEW检验, 以及联合Tajima's D、Fay & Wu's H和EW的DHEW检验, 与之前的DH检验相比, 这两种方法更不容易受到重组的影响, 在检测正选择时具有更高的统计学功效.

在此基础上, Grossman等人[58]于2010年提出的CMS(composite of multiple signals)进一步联合了更多的方法来检测自然选择: 包括Tajima's D, Fay & Wu's H, EW检验[57,59], F_{ST}, iHS和XP-EHH(the cross-population extended haplotype homozygosity), 还结合了两种作者提出的、基于频谱的检测方法: ΔDAF和ΔiHH. 因此这一方法考虑到了全部三类常用的自然选择检测信号: 长程单倍型(long-range haplotype), 稀有或高频突变相对过多以及相对较高的群体间遗传分歧. Simonson等人[60]使用CMS方法鉴别出了西藏人群中受到正选择的基因*EGLN1*和*PPARA*, 并发现了这些基因与高原人群血红蛋白表型之间的显著关联, 有助于理解人类如何适应严酷的高原环境.

1.3 群体历史与自然选择的联合估计

在面对真实情况时, 各个检验统计量对于群体历史等其他因素的敏感性不尽相同. 简单地联合多种方法来检测自然选择, 在群体曾受到瓶颈效应影响的情况下, 无法保证能够可靠地检测到自然选择[61], 需要进一步对群体历史进行分析.

一般来说, 不同于自然选择的局部作用方式, 群体历史事件会在全基因组的水平上影响DNA多态性. Nielsen等人[62]据此提出了一种新的方法: 将全基因组水平的突变频谱作为背景, 检测相关区域的突变频谱是否偏离了背景频谱. 然而该方法仅仅考虑了群体历史对于突变频谱均值的影响而忽略了群体历史有可能增加基因组突变频谱的方差, 因此在群体曾受到瓶颈效应影响的情况下, 依然会有较高的假阳性[63].

为了解决这一问题, Li和Stephan[64]提出了基于似然度(likelihood)的数据分析两步法: 利用全基因组(或部分基因组, 一般是非编码区)的SNP数据先估计群体历史, 再利用滑动窗口检验发生在染色体局部区域的自然选择. 这一方案囊括了群体历史对于突变频谱均值和方差两方面的影响, 在后续的大规模模拟验证中被证明可靠[63]并在许多研究中得到了普遍运用. 其中, 群体历史可以采用多种方法来估计, 如PSMC(the pairwise sequentially Markovian coalescent)[65], MSMC(the multiple sequentially Markovian coalescent)[66]、Stairway plot[67], 或者TNSFS(the total number of segregating sites as a function of sample size)[68]等方法. 但是Bank等人[55]也指出, 基于似然度的两步分析方法虽然尝试区分了群体历史和自然选择对基因组的影响, 仍存在一个不容忽视的问题: 如果不能正确估计群体历史, 那么后续的、对自然选择的检测结果也会存在偏差.

当所研究的群体历史较为复杂时，可能会难以进行两步法估计，似然度计算也相对棘手，或者在数学上难以求解．此时不需要两步法、不需要使用似然度计算的近似贝叶斯算法(approximate Bayesian computation, ABC)可能是更好的解决方案．该方法可以共同估计群体历史、背景选择(background selection, BGS)和自然选择，主要原理是利用贝叶斯定理，用已知参数的先验概率得到目标参数的后验分布．Bank等人[55]认为近似贝叶斯算法是以上共同估计方法中效率最高的．

1.4　溯祖树、祖先重组图与自然选择

近期发生的自然选择会改变溯祖事件(coalescent event)的产生速度，从而改变溯祖树(coalescent tree)使其与中性假设条件下的形态不同[69]．因此通过观察枝长不平衡的溯祖树，应当可以在特定的基因组区域检测近期发生的自然选择．中国科学院计算生物学研究所的Li[70]以此为依据，提出了基于溯祖树的MFDM (the maximum frequency of derived mutations)检验．使用模拟数据进行的分析表明，该方法能够衡量特定位点溯祖树的枝长不平衡程度，进而推断该位点是否经历了自然选择．这一过程不会受到群体大小变化的影响．如果使用特定的抽样方法，MFDM检验还能消除群体结构(population structure)带来的混淆．Li和Wiehe[71]进一步的研究工作发现，基于SNP数据和溯祖树的检测方法也可以扩展到微卫星(microsatellite)数据，并提出了衡量溯祖树拓扑结构的Ω统计量．通过使用真实序列的微卫星数据，对Ω统计量进行估计，就能够推测序列是否经历了选择扫荡．最近该研究组的Yang等人[72]进一步提出了基于D_u统计量的方法．使用该方法时，首先以有利突变事件发生的时间点将根据序列信息推断出的溯祖树划分为两个子树：群体历史事件对子树形态的影响相当，而自然选择会使两个子树的结构出现不平衡．通过计算两个子树的D_u值，比较差异，就可以在不受群体历史影响的情况下推断自然选择是否发生．

中国科学院计算生物学研究所和复旦大学的Wang等人[73]也基于条件溯祖树成功开发了SCCT(conditional coalescent tree)方法，根据有利突变事件将溯祖树划分为两个子树，分别统计和比较两个子树中的单倍型数量．在分别使用模拟数据和真实数据进行的分析中，与多种经典方法的比较均表明，该方法能够更快、更准确地找到正选择的候选基因．其中，在分析人群基因组数据时，SCCT方法成功定位了4个知名的、受正选择影响的基因: *ADH1B*, *MCM6*, *APOL1*和*HBB*．并且Disanto等人[74]从数学的角度深入探讨了Ω树的性质，以及衡量不平衡程度的具体方法．Ronen等人[75]延续了借助溯祖树定位有利突变的思想，提出了计算单倍型频率的HAF(the haplotype allele frequency)分值和应用HAF分值在选择扫荡影响的基因组数据中定位具体单倍型的PreCIOSS(predicting carriers of ongoing selective sweeps)算法．模拟数据和真实数据中的分析均表明，该方法能够较为可靠地确定携带有利突变的单倍型．在一项最新的研究中，Akbari等人[76]在HAF分值的基础上开发了iSAFE(integrated selection of allele favored by evolution)方法，该方法比SCCT方法更优，能够在较大基因组片段(~5 Mbp)中精确定位有利突变的SNP位点．

基于溯祖树的策略也可以用来检测其他类型的自然选择．Hunter-Zinck和Clark[77]开发的TSel方法能够在全基因组水平上检测正选择和平衡选择．与其他方法的比较分析表明，TSel方法在不同时间尺度和选择强度条件下都能取得较好的结果．

2　基于有监督学习的新方法

在过去的半个多世纪里，测序技术的突飞猛进促使群体基因组学数据经历了爆炸式增长．如何从海量数据中寻得规律、得出结论是目前研究者面临的一大挑战. Schrider和Kern[78]总结认为, 在这种情况下, 从单一概率模型或其近似进行参数估计的经典统计学方法可能是不够的, 应该考虑更强大的机器学习方法. 不同于经典统计学方法先构建模型、再将模型应用于感兴趣的数据，机器学习方法可以直接而自动地从海量数据中学得模型．相比于受制于“维度诅咒(the curse of dimensionality)”的传统方法，机器学习方法更能适应高维数据, 甚至更容易从中找到隐含的规律.

目前主流的机器学习方法可以分为四大类：无监督学习[79]、有监督学习[80]、半监督学习[81]和强化学习[82]．无监督学习算法擅长在缺乏先验知识(prior knowledge)的情况下发现数据集中的隐藏结构．在群体基因组学研究中已得到广泛应用的主成分分析

(principal component analysis, PCA)是典型的无监督学习, 它通常以非常高维的基因型矩阵作为输入, 然后生成一个维度较低的输出, 以揭示这些基因型如何聚类. 使用这种方法, 研究者能够评估未知个体或群体间的亲缘关系远近. Novembre等人[83]使用PCA分析了3192名欧洲人的SNP数据, 发现这些个体的聚类情况在很大程度上反映了个体所处的地理位置关系. 有监督学习算法需要使用训练数据集(training set)进行训练, 其中所有的数据都是有标签的数据(labeled data), 即每条输入数据都有对应的正确输出. 算法通过比较模型输出和正确输出, 发现错误、进行学习、对模型进行相应的改进. 随着算法不断迭代, 错误越来越少, 输出也越来越准确. 支持向量机(support vector machine, SVM), Boosting, 随机森林(random forest), 决策树(decision tree), 神经网络(neural network)都属于有监督学习算法.

在最理想的情况下, 研究者能够获得大量有标签的数据用于机器学习算法的训练, 但在面对具体问题时, 可能难以获得足够多的有标签数据. 此时可以考虑使用半监督学习. 半监督学习的训练数据只有少部分有标签(labeled), 而大部分没有, 其算法迭代过程与有监督学习类似. 强化学习通常用于机器人、游戏和导航领域. 此类学习算法有智能体(agent)、环境(environment)和行动(actions)三个主要组成部分, 智能体通过与环境交互、确定策略并采取行动, 以达到在给定时间内最大化预期奖励(reward)(如赢得游戏)的目标. Google旗下DeepMind公司开发的AlphaZero就用到了强化学习算法[84]. 相比于无监督学习和有监督学习, 半监督学习和强化学习目前鲜见于群体基因组学研究中. 后文将会主要介绍群体遗传学研究中对有监督学习的应用.

近年来, 逐渐有研究发现, 机器学习算法能够获得比传统统计学方法更好的表现. 但此类算法仍存在可能的缺陷. 例如, 在使用有监督学习中的支持向量机时, 需要人为选择适用于所研究问题的核函数(kernel function), 如果数据集很大, 核函数的选择也会相对困难, 选择不当则会带来结果的偏差. 相比于此类需要人为设定参数的有监督学习算法, 现代深度神经网络几乎不需要人为设定的先验知识(prior knowledge), 却会遇到训练好的模型难以解释的问题, 即无法确定该模型是否以期望的方式习得了数据背后的规律. 相关领域的研究者仍在努力理解神经网络的“黑箱(black-box)”, 例如, Zeiler和Fergus[85]提出的“反卷积(deconvolution)”概念就是解释卷积神经网络(convolutional neural network, CNN)的一种尝试. 全面考虑到机器学习算法的优点和可能的缺陷, 目前的趋势是综合多种算法来构建模型. AlphaZero不仅使用了强化学习, 还使用了卷积神经网络和蒙特卡洛搜索树(Monte Carlo tree search, MCTS)算法[84], 多种方法的综合使用有助于扬长避短, 更好地解决问题.

2.1　有监督学习与自然选择

在新近开始的一系列研究中, 使用有监督学习检测自然选择的方法获得了比传统统计学方法更好的表现. Pavlidis等人[86]使用支持向量机将两个经典统计量结合在一起, 发现该算法在检测选择扫荡时具有更高的统计学功效. 几乎同一时间, 中国科学院计算生物学研究所和维也纳大学的Lin等人[61]使用另外一类机器学习的方法Boosting, 将θ_W, θ_π, θ_H, Tajima's D, Fay & Wu's H以及iHH这6个经典统计量结合为一个特征向量, 用于算法的训练. 为准确区分自然选择和群体历史事件, 该方法使用了两个分类器(classifier): 首先使用第一个分类器预测输入数据是否受到自然选择, 再使用第二个分类器确定第一步检测到的是自然选择还是瓶颈效应(bottleneck). 根据Boosting分类器给出的权重高低, 还可以判断在不同的情形下, 不同统计量在检测自然选择时的相对重要程度. 研究发现, 在检测近期发生的自然选择时, iHH能提供更有用的信息; 而在检测历史较久远的自然选择时, θ_π更为有效; θ_W则较为擅长区分自然选择和瓶颈效应. 在一项新近的研究中, Schrider和Kern[87]使用随机森林的方法, 对自然选择进行了更精确的分类.

在群体历史和自然选择的联合估计方面, 有监督学习方法也取得了较好的效果. Sheehan和Song[88]提出的evoNet将近些年流行的深度神经网络引入群体基因组学研究中, 不仅能够推断群体大小的变化情况, 还能检测选择扫荡以及平衡选择影响的区域.

2.2　有监督学习在群体遗传学领域的进一步运用

使用有监督学习估计群体历史的方法同样获得了不错的表现. 近似贝叶斯算法是目前最为流行的群体历史估计方法, 但最近的研究发现, 使用有监督学

习方法有助于获得更优的结果. 使用经典近似贝叶斯算法估计群体历史时, 随着候选参数维度升高, 算法性能会显著下降. 因此Pudlo等人[89]开发了使用随机森林筛选群体历史模型的方法, 减少了所需的计算资源, 并取得了比标准近似贝叶斯算法更好的表现.

有监督学习方法还可以用于群体重组率ρ的估计. 中国科学院计算生物学研究所的Lin等人[90]使用Boosting方法, 以多个经典统计量作为解释变量, 以群体重组率ρ作为因变量, 训练算法进行回归分析, 估计群体重组率. 最终得到的Boosting算法模型赋予各统计量的权重, 能够在一定程度上反映各统计量对于群体重组率ρ估计的相对重要性. 测试发现, 该方法能够在几分钟内完成其他方法需要几天、甚至数年才能完成的分析任务. Gao等人[91]在此基础上, 针对全基因组测序数据开发了开源软件包FastEPRR. 该软件包在单核CPU上使用千人基因组计划单个人群的数据进行遗传重组率图谱绘制只需要3天时间, 比估计遗传重组率的经典软件Auton和McVean[92]开发的LDhat快30万倍且同样精确.

3　总结与展望

理解自然选择、群体结构和数量变化历史等力量如何共同塑造了基因组中的遗传变异模式是群体基因组学的重要任务. 经过一个多世纪的积累, 研究人员已经开发出了大量的方法, 但随着基因组学时代大量测序数据的爆炸式增长, 挑战更加严峻. 并且几种演化力量的相互作用也十分复杂, 例如, 自然选择的发生常常与群体历史事件(例如迁移、扩张或瓶颈效应)相伴. 计算所得结果可能提供的多种解释、统计学检验中可能产生的假阳性以及基因组数据中可能存在的系统偏差, 都使得群体基因组学检测自然选择和估计群体历史的整个过程成为一项艰巨的任务.

其实这一问题贯穿群体遗传学近期的发展史. 例如Enard等人[93]研究发现, 大名鼎鼎的语言基因*FOXP2*曾经受到过现代人特异的正选择作用并将有利突变(causal mutation)定位到了两个非同义突变上, 在Enard等人[94]后续的功能试验中, 将人源*FOXP2*敲入小鼠后, 小鼠也表现出了声学性状的改变. 然而多项对古DNA序列数据的分析表明, *FOXP2*的两个非同义突变在多种古人类中也存在[95,96], 不应该是驱动该次选择事件的有利突变[97], 与现代人类的语言形成无关[98]. 最近, Atkinson等人[99]重新对*FOXP2*及其附近的基因组区域进行了分析, 包括检验正选择信号、单倍型、频谱, 还进行了必要的实验验证. 结果发现, 没有证据可以证实*FOXP2*受到过现代人特异的正选择作用. 由此可见, 由于现有方法的局限性和生物进化历史的复杂性, 检测到的自然选择事件往往并不能代表复杂演化过程的全貌[100]. 然而, 理论群体遗传学的方法一旦在这一方面获得突破性进展, 对这一领域的促进作用将是极为明显的, 通过准确分析数十万或数百万年内进化的信息, 揭示出基因与环境的相互作用, 从而在进化生物学层次上研究基因的功能.

综上所述, 目前国际、国内检测自然选择的最新研究, 主要侧重于三个方向: 结合群体历史的信息、基于溯祖树和祖先重组图的方法以及基于有监督学习的方法. 研究人员能在多大程度上从不断增长的数据集中获取有意义的结论取决于其是否能够使用多有效的方法. 目前, 研究人员已经使用各种传统方法在群体历史、人类演化以及作物驯化等方面取得了众多有意义的科学成果. 随着新方法的不断涌现, 在可以预见的未来, 应用这些方法进行的研究有望催生更多的科学成果. 一方面, 传统的统计学方法仍然实用而且通常可以获得比较可靠的结果, 预计在将来仍将占据主导地位; 另一方面, 新型有监督学习方法可以为困扰研究人员的旧问题提供新的解决方案. 工业界和生物信息学领域的有监督学习方法经常受限于能否获得大量的、高质量有标签数据, 然而借助于模拟软件[75,101], 群体基因组学家几乎可以无限量地获得所需要的有标签数据. 得益于这一巨大优势以及机器学习擅长处理高维数据的特点, 有监督学习方法将可以很大程度上助力群体基因组学研究.

参考文献

1　Darwin C. On The Origin of Species By Means of Natural Selection, or The Preservation of Favoured Races in The Struggle For Life. London: John Murray, 1859. 126–127

2 Huxley J. Evolution: The Modern Synthesis. London: George Allen and Unwin, 1942. 1–45
3 Crow J F. Population genetics history: a personal view. Annu Rev Genet, 1987, 21: 1–22
4 Kimura M. Evolutionary rate at the molecular level. Nature, 1968, 217: 624–626
5 Hughes A L. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity, 2007, 99: 364–373
6 Voight B F, Kudaravalli S, Wen X, et al. A map of recent positive selection in the human genome. PLoS Biol, 2006, 4: e72
7 Consortium T I H. A haplotype map of the human genome. Nature, 2005, 437: 1299–1320
8 Gibbs R A, Boerwinkle E, Doddapaneni H, et al. A global reference for human genetic variation. Nature, 2015, 526: 68–74
9 Sudmant P H, Rausch T, Gardner E J, et al. An integrated map of structural variation in 2,504 human genomes. Nature, 2015, 526: 75–81
10 Walter K, Min J L, Huang J, et al. The UK10K project identifies rare variants in health and disease. Nature, 2015, 526: 82–90
11 Geihs M, Yan Y, Walter K, et al. An interactive genome browser of association results from the UK10K cohorts project. Bioinformatics, 2015, 31: 4029–4031
12 Hoban S, Bertorelle G, Gaggiotti O E. Computer simulations: tools for population and evolutionary genetics. Nat Rev Genet, 2012, 13: 110–122
13 Ke Y, Su B, Song X, et al. African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. Science, 2001, 292: 1151–1153
14 He Y X, Qi X B, Ouzhuluobu X, et al. Blunted nitric oxide regulation in Tibetans under high-altitude hypoxia. Natl Sci Rev, 2018, 5: 516–529
15 Marciniak S, Perry G H. Harnessing ancient genomes to study the history of human adaptation. Nat Rev Genet, 2017, 18: 659–674
16 Wang G D, Shao X J, Bai B, et al. Structural variation during dog domestication: insights from grey wolf and dhole genomes. Natl Sci Rev, 2019, doi: 10.1093/nsr/nwy076
17 Ling S, Hu Z, Yang Z, et al. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. Proc Natl Acad Sci USA, 2015, 112: E6496–E6505
18 Wu C I, Wang H Y, Ling S, et al. The ecology and evolution of cancer: the ultra-microevolutionary process. Annu Rev Genet, 2016, 50: 347–369
19 Wang H Y, Chen Y X, Tong D, et al. Is the evolution in tumors Darwinian or non-Darwinian? Natl Sci Rev, 2018, 5: 15–17
20 Schneider A, Souvorov A, Sabath N, et al. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol, 2009, 1: 114–118
21 Wright S. Evolution in Mendelian populations. Genetics, 1931, 16: 97–159
22 Wright S. Breeding structure of populations in relation to speciation. Am Natist, 1940, 74: 232–248
23 Watterson G A. On the number of segregating sites in genetical models without recombination. Theor Populat Biol, 1975, 7: 256–276
24 Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 1989, 123: 585–595
25 Fu Y X, Li W H. Statistical tests of neutrality of mutations. Genetics, 1993, 133: 693–709
26 Fay J C, Wu C I. Hitchhiking under positive Darwinian selection. Genetics, 2000, 155: 1405–1413
27 Zeng K, Fu Y X, Shi S, et al. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics, 2006, 174: 1431–1439
28 Fu Y X. A phylogenetic estimator of effective population size or mutation rate. Genetics, 1994, 136: 685–692
29 Fu Y X, Li W H. Maximum likelihood estimation of population parameters. Genetics, 1993, 134: 1261–1270
30 Wright S. The genetical structure of populations. Ann Eugen, 1949, 15: 323–354
31 Weir B S, Cockerham C C. Estimating *F*-statistics for the analysis of population structure. Evolution, 1984, 38: 1358–1370
32 Hudson R R, Kaplan N L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics, 1985, 111: 147–164
33 Nielsen R. Molecular signatures of natural selection. Annu Rev Genet, 2005, 39: 197–218
34 Vitti J J, Grossman S R, Sabeti P C. Detecting natural selection in genomic data. Annu Rev Genet, 2013, 47: 97–120
35 Zhou Q, Wang W. Detecting natural selection at the DNA level (in Chinese). Zool Res, 2004, 25: 73–80 [周琦, 王文. DNA水平自然选择作用的检测. 动物学研究, 2004, 25: 73–80]
36 Lin K, Li H P. Advances in detecting positive selection on genome (in Chinese). Hereditas, 2009, 31: 896–902 [林栲, 李海鹏. DNA水平上检测正选择方法的研究进展. 遗传, 2009, 31: 896–902]
37 Chen H, Hey J, Slatkin M. A hidden Markov model for investigating recent positive selection through haplotype structure. Theor Popul Biol,

2015, 99: 18–30
38 Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res, 2010, 20: 393–402
39 Yi X, Liang Y, Huerta-Sanchez E, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science, 2010, 329: 75–78
40 Akey J M. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res, 2009, 19: 711–722
41 Li W H, Wu C I, Luo C C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol, 1985, 2: 150-–174
42 Hurst L D. The K_a/K_s ratio: diagnosing the form of sequence evolution. Trends Genets, 2002, 18: 486–487
43 McDonald J H, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature, 1991, 351: 652–654
44 Pollard K S, Salama S R, King B, et al. Forces shaping the fastest evolving regions in the human genome. PLoS Genet, 2006, 2: e168
45 Pollard K S, Salama S R, Lambert N, et al. An RNA gene expressed during cortical development evolved rapidly in humans. Nature, 2006, 443: 167–172
46 Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics, 2002, 160: 765–777
47 Sabeti P C, Reich D E, Higgins J M, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature, 2002, 419: 832–837
48 Wang E T, Kodama G, Baldi P, et al. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci USA, 2006, 103: 135–140
49 Lewontin R C, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics, 1973, 74: 175–195
50 Shriver M D, Kennedy G C, Parra E J, et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genom, 2004, 1: 274–286
51 Jaillon O, Aury J M, Brunet F, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature, 2004, 431: 946–957
52 Hufford M B, Xu X, van Heerwaarden J, et al. Comparative population genomics of maize domestication and improvement. Nat Genet, 2012, 44: 808–811
53 Zhou Z, Jiang Y, Wang Z, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol, 2015, 33: 408–414
54 Li J, Li H, Jakobsson M, et al. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol Ecol, 2012, 21: 28–44
55 Bank C, Ewing G B, Ferrer-Admettla A, et al. Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genets, 2014, 30: 540–546
56 Zeng K, Shi S, Wu C I. Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol, 2007, 24: 1898–1908
57 Watterson G A. The homozygosity test of neutrality. Genetics, 1978, 88: 405–417
58 Grossman S R, Shlyakhter I, Shylakhter I, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science, 2010, 327: 883–886
59 Ewens W J. The sampling theory of selectively neutral alleles. Theor Popul Biol, 1972, 3: 87–112
60 Simonson T S, Yang Y, Huff C D, et al. Genetic evidence for high-altitude adaptation in Tibet. Science, 2010, 329: 72–75
61 Lin K, Li H, Schlötterer C, et al. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics, 2011, 187: 229–244
62 Nielsen R, Williamson S, Kim Y, et al. Genomic scans for selective sweeps using SNP data. Genome Res, 2005, 15: 1566–1575
63 Pavlidis P, Hutter S, Stephan W. A population genomic approach to map recent positive selection in model species. Mol Ecol, 2008, 17: 3585–3598
64 Li H, Stephan W. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet, 2006, 2: 1580–1589
65 Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature, 2011, 475: 493–496
66 Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet, 2014, 46: 919–925
67 Liu X, Fu Y X. Exploring population size changes using SNP frequency spectra. Nat Genet, 2015, 47: 555–559
68 Chen H, Hey J, Chen K. Inferring very recent population growth rate from population-scale sequencing data: using a large-sample coalescent

estimator. Mol Biol Evol, 2015, 32: 2996–3011
69 Kaplan N L, Hudson R R, Langley C H. The hitchhiking effect revisited. Genetics, 1989, 123: 887–899
70 Li H. A new test for detecting recent positive selection that is free from the confounding impacts of demography. Mol Biol Evol, 2011, 28: 365–375
71 Li H, Wiehe T. Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. PLoS Comput Biol, 2013, 9: e1003060
72 Yang Z, Li J, Wiehe T, et al. Detecting recent positive selection with a single locus test bipartitioning the coalescent tree. Genetics, 2018, 208: 791–805
73 Wang M, Huang X, Li R, et al. Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree. Mol Biol Evol, 2014, 31: 3068–3080
74 Disanto F, Schlizio A, Wiehe T. Yule-generated trees constrained by node imbalance. Math Biosci, 2013, 246: 139–147
75 Ronen R, Tesler G, Akbari A, et al. Predicting carriers of ongoing selective sweeps without knowledge of the favored allele. PLoS Genet, 2015, 11: e1005527
76 Akbari A, Vitti J J, Iranmehr A, et al. Identifying the favored mutation in a positive selective sweep. Nat Meth, 2018, 15: 279–282
77 Hunter-Zinck H, Clark A G. Aberrant time to most recent common ancestor as a signature of natural selection. Mol Biol Evol, 2015, 32: 2784–2797
78 Schrider D R, Kern A D. Supervised machine learning for population genetics: a new paradigm. Trends Genets, 2018, 34: 301–312
79 Ghahramani Z. Unsupervised Learning. Advanced Lectures On Machine Learning. Heidelberg: Springer, 2003. 72–112
80 Kotsiantis S B. Supervised machine learning: a review of classification techniques. Informatica (lithuanian Academy of Sciences), 2007, 31: 249–268
81 Zhu X J, Goldberg A B. Introduction To Semi-supervised Learning. San Rafael: Morgan & Claypool, 2009. 9–20
82 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
83 Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. Nature, 2008, 456: 98–101
84 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550: 354–359
85 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. Lect Notes Comput Sc, 2014, 8689: 818–833
86 Pavlidis P, Jensen J D, Stephan W. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics, 2010, 185: 907–922
87 Schrider D R, Kern A D. S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet, 2016, 12: e1005928
88 Sheehan S, Song Y S. Deep learning for population genetic inference. PLoS Comput Biol, 2016, 12: e1004845
89 Pudlo P, Marin J M, Estoup A, et al. Reliable ABC model choice via random forests. Bioinformatics, 2016, 32: 859–866
90 Lin K, Futschik A, Li H. A fast estimate for the population recombination rate based on regression. Genetics, 2013, 194: 473–484
91 Gao F, Ming C, Hu W, et al. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. G3, 2016, 6: 1563–1571
92 Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome Res, 2007, 17: 1219–1227
93 Enard W, Przeworski M, Fisher S E, et al. Molecular evolution of *FOXP2*, a gene involved in speech and language. Nature, 2002, 418: 869–872
94 Enard W, Gehre S, Hammerschmidt K, et al. A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell, 2009, 137: 961–971
95 Krause J, Lalueza-Fox C, Orlando L, et al. The derived *FOXP2* variant of modern humans was shared with Neandertals. Curr Biol, 2007, 17: 1908–1912
96 Maricic T, Günther V, Georgiev O, et al. A recent evolutionary change affects a regulatory element in the human *FOXP2* gene. Mol Biol Evol, 2013, 30: 844–852
97 Coop G, Bullaughey K, Luca F, et al. The timing of selection at the human *FOXP2* gene. Mol Biol Evol, 2008, 25: 1257–1259
98 Preuss T M. Human brain evolution: from gene discovery to phenotype discovery. Proc Natl Acad Sci USA, 2012, 109: 10709–10716
99 Atkinson E G, Audesse A J, Palacios J A, et al. No evidence for recent selection at *FOXP2* among diverse human populations. Cell, 2018, 174: 1424–1435.e15
100 Xiang-Yu J, Yang Z, Tang K, et al. Revisiting the false positive rate in detecting recent positive selection. Quant Biol, 2016, 4: 207–216

101 Gao F, Li H P. Application of computer simulators in population genetics (in Chinese). Hereditas, 2016, 38: 707–717 [高峰, 李海鹏. 群体遗传学模拟软件应用现状. 遗传, 2016, 38: 707–717]

Population genomics: From classical statistics to supervised learning

SHI Yi[1,3] & LI HaiPeng[1,2]

1 Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;
2 Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China;
3 University of Chinese Academy of Sciences, Beijing 100049, China

It is essential to understand how the patterns of genetic variation in organisms have been shaped by different evolutionary forces, such as mutation, natural selection, genetic drift, population structure, and population size change. In recent years, with the rapid innovation of next-generation sequencing technology, we are facing the new era of population genomics. The relevant population genomics methods can be classified as classical statistics and supervised learning. The classical statistics methods include many popular ones for detecting natural selection and inferring the parameters of demography, which are based on single or multiple combined statistics. The supervised learning methods may promise a new paradigm to make sense of large datasets in the genomic era. Here a brief introduction was first given on the important theory in population genomics. Then we overviewed the recent research progress in population genomics and shared our perspectives on its future development.

population genomics, natural selection, recombination rate, classical statistics, supervised learning

doi: 10.1360/N052018-00207